Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

DeepSVP: Integration of genotype and phenotype for structural variant prioritization using deep learning

Azza Althagafi [1,2], Lamia Alsubaie [3], Nagarajan Kathiresan [4], Katsuhiko Mineta [1], Taghrid Aloraini [3], Fuad Almutairi [7,8], Majid Alfadhel [5,7,8], Takashi Gojobori [9], Ahmad Alfares [3,5,6], and Robert Hoehndorf[1,*]

[1] Computational Bioscience Research Center, Computer, Electrical and Mathematical Sciences & Engineering Division, King Abdullah University of Science and Technology, 4700 King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia; [2] Computer Science Department, College of Computers and Information Technology, Taif University, Taif 26571, Saudi Arabia; [3] Department of Pathology and Laboratory Medicine, King Abdulaziz Medical City, Riyadh, Saudi Arabia; [4] Supercomputing Core Lab, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia; [5] King Abdullah International Medical Research Center, Riyadh, Saudi Arabia; [6] Department of Pediatrics, College of Medicine, Qassim University, Qassim, Saudi Arabia; [7] Division of Genetics, Department of Pediatrics, King Abdulaziz Medical City, Riyadh, Saudi Arabia; [8] King Saud bin Abdulaziz University for Health Sciences, King Abdulaziz Medical City, Riyadh, Saudi Arabia; [9] Biological and Environmental Science and Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Structural genomic variants account for much of human variability and are involved in several diseases. Structural variants are complex and may affect coding regions of multiple genes, or affect the functions of genomic regions in different ways from single nucleotide variants. Interpreting the phenotypic consequences of structural variants relies on information about gene functions, haploinsufficiency or triplosensitivity, and other genomic features. Phenotype-based methods to identifying variants that are involved in genetic diseases combine molecular features with prior knowledge about the phenotypic consequences of altering gene functions. While phenotype-based methods have been applied successfully to single nucleotide variants, as well as short insertions and deletions, the complexity of structural variants makes it more challenging to link them to phenotypes. Furthermore, structural variants can affect a large number of coding regions, and phenotype information may not be available for all of them.

Results: We developed DeepSVP, a computational method to prioritize structural variants involved in genetic diseases by combining genomic information with information about gene functions. We incorporate phenotypes linked to genes, functions of gene products, gene expression in individual celltypes, and anatomical sites of expression, and systematically relate them to their phenotypic consequences through ontologies and machine learning. DeepSVP significantly improves the success rate of finding causative variants in several benchmarks and can identify novel pathogenic structural variants in consanguineous families.

Availability: `https://github.com/bio-ontology-research-group/DeepSVP`

Contact: robert.hoehndorf@kaust.edu.sa

1 Introduction

Structural genomic variants are genomic variants that affect more than 50 base pairs and include copy number variants, insertions, and deletions (Eichler, 2019). Many structural variants are implicated in heritable diseases (Sudmant *et al.*, 2015). While there have been several efforts to predict and prioritize pathogenic genomic variants (Eilbeck *et al.*, 2017), predicting the functional impact of structural variants discovered through genome sequencing studies remains challenging due to the diversity of variant size and type; structural variants may cover multiple coding and non-coding regions, overlap several genes, and are affected by haploinsufficiency and triplosensitivity (Kidd *et al.*, 2008).

Methods for predicting the pathogenicity of genomic variants may be based on their impact on protein structure, measures of sequence conservation, or function (Eilbeck *et al.*, 2017). However, due to the complexity of the structural variants, including the variant size, type, and overlap with multiple genes, designing methods that determine structural variant pathogenicity is more challenging. Several efforts for analyzing the clinical impact of structural variants have focused on well-matched cases and controls. For instance, by evaluating the loci and the respective pathways that may be impacted by a structural variant at these loci, it became possible to define novel genes involved in complex disorders such as autism (Pinto *et al.*, 2010) or immune-related disorders (Rossin *et al.*, 2011). While there are several methods to identify disease-associated variants in cohorts, it is more challenging to discover disease-associated variants that exist in a single sample or pedigree, in particular in rare Mendelian disorders (Sanchis-Juan *et al.*, 2018).

Methods that evaluate the functional consequence of structural variants in individual genomes use different strategies. Several approaches include genomic information, such as variant length, haploinsufficiency measures, or GC content, to separate pathogenic from benign structural variants (Hehir-Kwa *et al.*, 2010; Sharo *et al.*, 2020). Furthermore, the predicted pathogenicity of deleterious single nucleotide variants within a structural variant can be used to estimate pathogenicity of structural variants (Ganel *et al.*, 2017). Additionally, phenotypes associated with a loss of function in single genes has also been used for prioritizing structural variants (Doelken *et al.*, 2012; Köhler *et al.*, 2014).

Phenotype-driven variant prioritization methods aim to link variants to the phenotypes observed in individuals using prior knowledge (Eilbeck *et al.*, 2017). Commonly, the link is established using a similarity measure between phenotypes associated with a variant or gene and the phenotypes observed in a patient (Smedley *et al.*, 2015). Phenotype-based methods are successful in finding disease-associated variants (Shefchek *et al.*, 2020) but suffer from the limited information about variant– or gene–phenotype associations. One way to overcome this limitation is to utilize phenotypes observed in model organisms and link them to human phenotypes (Smedley *et al.*, 2013; Shefchek *et al.*, 2020); however, even when including phenotypes from model organisms a large portion of human protein-coding genes remain without associations, thereby limiting the success of phenotype-based methods to variants or genes that have previously been studied either in humans or animal models, or relying on guilt-by-associations approaches in which information about phenotypes is propagated through associations such as interaction networks (Smedley *et al.*, 2014).

Several deep learning methods are now available that can predict phenotypes (Zhou *et al.*, 2019; Kulmanov and Hoehndorf, 2020) or associate phenotypes with different types of information available for genes, including functions of gene products and site of expression (Chen *et al.*, 2020; Smaili *et al.*, 2019). These methods use machine learning to relate information through background knowledge contained in ontologies, and can accurately identify phenotype-associated genes without prior knowledge about phenotypes, often significantly improving over the use of semantic similarity measures (Kulmanov *et al.*, 2020). A limitation of these methods is that they are usually transductive instead of inductive (Kulmanov *et al.*, 2020), i.e., the diseases or disorders for which associated genes are predicted should already be available at the time of training the model. As these methods require information about disease-associated phenotypes during training, they will therefore not generalize to entirely new cases, thereby limiting their application in identifying phenotype-associated genomic variants.

We developed a machine learning method that predicts whether a structural variant (duplication or deletion) is pathogenic and involved in the development of specific phenotypes. Our method combines genomic information and clinical phenotypes, and leverages a large amount of background knowledge from human and animal models; for this purpose, we extend an ontology-based deep learning method to allow inductive inference. We demonstrate that our method improves over the state of the art in detecting pathogenic deletions or duplications. We further apply our method to the diagnosis of a family with congenital disease involving infantile spasms and seizures for which previous analysis of single nucleotide variants in whole exome and whole genome sequencing data found no associated variant. We make DeepSVP freely available as a Python package at `https://github.com/bio-ontology-research-group/DeepSVP`.

2 Materials and Methods

2.1 Data sources and ontologies

We use as training and testing dataset the set of pathogenic and benign structural variants (SVs) aligned to the human reference genome GRCh38 obtained from the database of genomic structural variation (dbVar) (Griffith and Griffith, 2004) downloaded on 8th Feb 2020. The dataset contains 14,197 (10,401 deletions, 3,796 duplications) pathogenic or likely pathogenic structural variants and 4,477 variants associated with one or more diseases (3,737 deletions, 586 duplications), as well as 25,890 (13,742 deletions, 12,148 duplications) benign or likely benign structural variants.

For each pathogenic structural variant, we defined variant–disease pairs with associated diseases from Online Mendelian Inheritance in Men (OMIM) database (Amberger *et al.*, 2011). There are 3,805 structural variants linked with one or more than one OMIM disease; if a variant is associated with n OMIM diseases, we generate n variant–disease pairs. As a result, we obtained 5,907 causative pathogenic variant–disease pairs. As negative training pairs we selected both benign and pathogenic variants and associate them with a randomly selected disease; we include pathogenic variants in the negative training pairs to simulate the case where variants may be pathogenic but not associated with the phenotypes observed in a patient (i.e., with a different phenotypes). After this step we obtained 36,041 negative (not causative) variant–disease pairs .

We use functional and phenotypic characteristics for genes in the human genome, in particular the phenotypes associated with human genes in the Human Phenotype Ontology (HPO) database (Köhler *et al.*, 2019), the phenotypes associated with mouse orthologs in the Mouse Genome Informatics (MGI) database (Bult *et al.*, 2019), the functions of gene products from UniProt (UniProt Consortium, 2019), gene expression in human celltypes (Tabula Muris Consortium *et al.*, 2018) and the anatomical site of expression from the GTEx tissue expression database (GTEx Consortium, 2015). These annotations are characterized using the HPO (Köhler *et al.*, 2019), Mammalian Phenotype Ontology (MP) (Bult *et al.*, 2019), Gene Ontology (GO) (Gene Ontology Consortium, 2019), Cell

Ontology (CL) (Diehl *et al.*, 2016), and the UBERON ontology (Mungall *et al.*, 2012).

We downloaded phenotypes from the HPO database on 16 July 2020 and obtain 169,281 phenotype associations for 4,315 human genes. We downloaded phenotype associations from MGI on 20 March 2020 and obtained 228,214 associations between 13,529 mouse genes and classes in MP. We identify the human ortholog for mouse genes using the `HMD_HumanPhenotype.rpt` orthology file from MGI and identify 10,951 human orthologs for the 13,529 mouse genes resulting in 168,550 associations between human genes and MP classes. We obtain the GO annotations for human gene products from the GO Annotation database (Huntley *et al.*, 2014) on 20 March 2020 for 18,495 gene products with 495,719 annotations in total. We filtered out all the GO annotations with the evidence code indicating that the annotation was inferred from electronic annotation (IEA), or no biological data is available (ND). We map the UniProt accessions of the gene products to Entrez gene identifiers using the mappings provided by the Entrez database (Maglott *et al.*, 2010), and obtain 17,786 genes which have GO annotations for their gene product resulting in 208,630 associations between genes and GO classes. For the anatomical location of gene expression, we downloaded the GTEx Tissue Expression Profiles from the Gene Expression Atlas (Papatheodorou *et al.*, 2020) which identifies gene expression across 53 tissues; 20,538 genes have an expression above the 4.0 threshold which was previously determined to be useful for predicting disease associations (Chen *et al.*, 2020) resulting in 585,765 associations between genes and UBERON classes. We represent the tissues with their UBERON classes, excluding the tissues *transformed skin fibroblast* and *EBV-transformed lymphocyte* as these are not found in UBERON. For the cell type, we downloaded single-cell RNAseq data from the Tabula Muris project (Tabula Muris Consortium *et al.*, 2018) in which genes are annotated with the CL. From this dataset, we obtain 6,559 human genes which have CL annotations, and 17,149 associations between genes and one or more classes from CL.

We use the combined PhenomeNET ontology (Rodríguez-García *et al.*, 2017) downloaded on 6 October, 2020, from the AberOWL ontology repository, as our phenotype ontology. PhenomeNET combines the phenotypes of human and other model organism as well as UBERON, GO and CL, and allows them to be compared.

2.2 Variant-based features

We annotate structural variants with a set of genomic features using public databases. We use AnnotSV v2.3 (Geoffroy *et al.*, 2018), which uses data from multiple external databases to annotate and rank SVs based on the overlapping regions of the variants with known pathogenic variants in dbVar (Griffith and Griffith, 2004), the Database of Genomic Variants (DGV) (MacDonald *et al.*, 2014), and disease-associated genes from OMIM. For each variant, AnnotSV generates annotations based on the variant length and the genes with which the variant overlaps (choosing among Refseq (O'Leary *et al.*, 2016) gene annotations). Furthermore, AnnotSV reports the list of promoters with which the variant overlaps. From the annotations provided by AnnotSV, we use the variant length, variant type, GC content around the variant's breakpoints (GCcontent_left, GCcontent_right), and the number of promoters and genes affected by the variant as features.

We further use AnnotSV to obtain information about genes with which a structural variant overlaps: the length of the Coding DNA Sequence (CDS), transcript length (tx length), haploinsufficiency ranks collected from the Deciphering Developmental Disorders (DDD) study (Firth and Wright, 2011), haploinsufficiency (HI_DDDpercent) and triplosensitivity estimates from ClinGen (Rehm *et al.*, 2015), gene intolerance annotations from ExAC (Karczewski *et al.*, 2017) including six annotations for synonymous variants (synZExAC), missense variants (misZExAC), loss of function variants (pLIExAC), deleation (delZExAC), duplications (dupZExAC), and CNV intolerance (cnvZExAC). Supplementary Table 1 summarizes all the features used in our predictive model.

While not used as a feature of our prediction model, we also use AnnotSV to identify the allele frequency of variants using the 1,000 genomes allele frequency (Sudmant *et al.*, 2015) and allele frequency from gnomAD (Collins *et al.*, 2020). We use this information to filter out common variants before applying our prediction.

2.3 Embedding ontology-based features

We use the DL2Vec (Chen *et al.*, 2020) method to encode ontology-based features associated with genes in a low-dimensional feature vector. DL2Vec "embeds" ontologies and their annotations in a real-valued vector space. For this purpose, DL2Vec first generates a graph $G = (V, E)$ from the ontology axioms in which nodes V represent classes or entities annotated with ontology classes, and edges E represent axioms that hold between these classes (Chen *et al.*, 2020). DL2Vec then explores the graph using random walks, and generates embeddings from these walks using Word2Vec.

We apply DL2Vec to the annotations and ontologies for the genes associated with human phenotypes (HPO), functions and cellular locations (GO), the phenotypes of their mouse orthologs (MP), expression in cell types (CL), and anatomical site of expression (UBERON). As each feature is available for a different number of genes, we generate embedding for each kind of feature separately. As parameters for DL2Vec, we use 100 random walks with a walk length of 25; we use the Word2Vec (Mikolov *et al.*, 2013) skip-gram model to generate the embeddings from the walks with 10 window size, 1 as the minimum count value, and an embedding size of 100. We train the skip-gram model for 20 epochs. As a result, we obtain a real-valued feature vector for each gene (and ontology class) of size 100.

Our model takes a set of phenotypes, encoded using the HPO, as input to represent patient phenotypes. We also encode these phenotypes using DL2Vec to generate an embedding of the patient phenotypes. As there will not likely be a representation of an entity associated with all phenotypes used as input to our method, we first update the graph used by DL2Vec to add a new node for the patient; this new node is associated with all phenotypes used as input. We then generate a new embedding for this node by performing random walks and then updating the pre-trained skip-gram model using these walks. This approach allows the skip-gram model to generate an embedding for a new patient (specified entirely by a set of phenotypes) while considering the full DL2Vec graph generated from the ontology.

2.4 Estimating variant pathogenicity by supervised prediction

We hypothesize that phenotypes we observe in patients resulting from a SV correlate with phenotypes observed when altering one or more of the genes affected by the variant and, therefore, that phenotypes associated with genes can provide additional information to predict potential causative SVs (Doelken *et al.*, 2012). Consequently, we develop a machine learning model to distinguish between causative and non-causative variants based on features derived from an individual samples and background knowledge.

We build a machine learning model that ranks SVs depending on their predicted pathogenicity and the relations between genes affected by the SV and the phenotype observed in affected individuals. Additionally, our approach consider several genomic features of each SVs, such as the coding sequence length overlapping with the SV, GC content, and haploinsufficiency and triplosensitivity scores to measure the dosage-sensitivity for genes/regions. Using the VCF file of the patient and

HPO-encoded set of phenotypes, all features are generated by annotating the VCF file. Figure 1 presents a high-level summary over our model. The DeepSVP prediction model consists of two parts: a phenotype prediction model based on matching the gene features with the phenotypes observed in the affected individuals; and a combined prediction model based on the phenotype prediction scores and the genomic features of the variant.

2.4.1 Phenotype prediction model

We apply DL2Vec (Chen *et al.*, 2020) to generate the feature representation for the patient phenotypes and genes. DL2Vec learns the "representation" for phenotypes and genes based on their annotations to ontology classes. The inputs to DL2Vec are associations of entities with ontology classes and the outputs are vectors (embeddings) of these entities. DL2Vec utilizes the axioms in ontologies to construct a graph representing phenotypes and their interrelations. DeepSVP incorporates biological background knowledge about the relation between phenotypes resulting from a loss of function in mouse/human genes, gene functions as defined using the Gene Ontology, as well as the celltype and anatomical site of gene expression.

The phenotype model takes two vectors v_1 and v_2 as input, representing the embedding for the patient's phenotypes and the embedding for a gene, respectively. The embeddings are used as input for two neural network models ν_1 and ν_2. We then calculate the inner product for $\nu_1(v_1)$ and $\nu_2(v_2)$ and apply a sigmoid activation function to generate a prediction score, between the embedding for the phenotypes v_1, and gene v_2. We use binary cross-entropy as a loss function to train our model.

Each neural network ν_1 and ν_2 consists of two hidden layers, in which the first layer with 256 units, and the second layer with 50 units. After each layer, we use dropout with a rate of 20% , followed by a Leaky Rectified Linear Unit (LeakyReLU) (Maas *et al.*, 2013) activation function. We use the Adam optimizer (Kingma and Ba, 2014) to optimize the model parameters. We develop five different models using DL2Vec embeddings based on different feature types: functions of gene products (GO), mouse model phenotypes (MP), human phenotype (HP), celltype (CL), and site of expression (UBERON). For each set of phenotypes (characterizing a disorder, or the clinical phenotypes observed in an individual), and for each prediction model, we rank each gene based on the DL2Vec prediction score, from smallest to highest, and represent the association between them by their m-quantile in this distribution (Breckling and Chambers, 1988) (Supplementary Figure 1 shows the distribution of the normalized quantile scores). This normalization and ranking aims to make prediction scores comparable across sets of phenotypes (Cornish *et al.*, 2018). We use the quantile as one of the features of the combined prediction models.

2.4.2 Combined prediction model

The combined prediction model uses the variant features and the phenotype-based scores produced by the DL2Vec-based predictions; the model is an artificial deep neural network model that uses genomic features derived from a variant as input together with the prediction score generated from the phenotype prediction model. We trained a separate model for each ontology dataset and aggregation type, either the maximum or average features scores for the genes within the variant region. The features used by the combined model are listed in the Supplementary Table 1.

Given a structural variant τ affecting regions that contain genes $G_1, ..., G_n$, we obtain the phenotype prediction score $\phi(G_i)$ for the genes $G_1, ..., G_n$ using each of the phenotype-based prediction models. We transform these scores into a feature for the variant τ using either the maximum or average of all the gene scores, i.e., either $\phi_{max}(\tau) = \max_{1 \leq i \leq n} \phi(G_i)$ or $\phi_{avg}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \phi(G_i)$. We normalize all the features using z-score normalization, in which the values for and feature F are scaled based on the mean, and standard deviation of F. The value v_i of F is normalized to $v_i^{'}$ by:

$$v_i^{'} = \frac{v_i - \mu_F}{\sigma_F} \tag{1}$$

where $v_i^{'}$ is z-score normalized one values, v_i is the i-th value for the feature A, μ_F is the mean, and σ_F the standard deviation, for feature F. We use the same mean and standard deviation to normalize the testing set.

We use 22 features for variants (8 features for the variant and 9 derived from the genes overlapping the variant) as well as 5 features from ontology embeddings. Some features are missing for some variants; to account for missing values, we use imputation. We imputed missing values by assigning them a zero value and additionally created indicator variables with a value set to 1 if the corresponding variant is missing and 0 otherwise. We use one-hot encoding to represent the categorical features with an "undefined" category for missing categorical annotations. We provide an analysis of feature importance and the correlation between features as Supplementary Materials.

2.4.3 Training and testing

Our prioritization algorithm relies on finding the "causal" variants, which are (a) pathogenic and (b) involved in developing a set of phenotypes observed in a patient. We investigated the performance of our machine learning classifier to predict the causative structural variant based on genomic characteristics of the variant, the functions and phenotypes that are likely altered by the variant, and the phenotypes observed in the individual in which the variant was detected. In our experiments, we considered as positive instances all the causative variants in our training set with the disease phenotypes for which they are causative. We separate the positive instances from two types of negatives instances, the benign variants which do not change any protein function and are not implicated in a disease, and pathogenic (but non-causative) variants that may be pathogenic but are not related to the disease phenotypes observed in a patient (but potentially related to another set of phenotypes).

We randomly split our dataset by diseases and associated variants into 85% for training and 15% for testing, where the set of diseases in training are different from the testing (to reflect the application of our method to entirely new diseases that have not been used for training). We use 15% of the training set as a validation set. We used the training and validation sets to train and tune model hyperparameters and select the best models, while the test set for reporting the evaluation metrics.

The number of negative samples (representing the set of variant-phenotype pairs that are either benign or the pathogenic but not causative variants) is significantly higher than the number of positives samples (representing the causative variants of a particular disease). For that reason, we randomly sample 5 times from the negative set a number of instances equal to the positive instances and use these as negatives for training.

We build a separate model for each ontology dataset and aggregation operation (either maximum or average). The architecture of the prediction model was derived by hyperparameter optimization using Scikit-Optimize with Bayesian optimization (`skopt`) repeated 15 times. We tuned for the following set of hyperparameters for the model: learning-rate $r \in [1e-6, 1e-2]$ with logarithmic transformation, and dropout $r \in [0.1, 0.5]$; number of layers $l \in [2, 6]$, and the number of nodes for each of the dense layers $n \in [50, 512]$; activation functions $a \in \{relu, sigmoid, selu\}$. Following our experiments, the optimal parameters (learning rate, number of dense layers, number of dense nodes, dropout, activation) for each model were as follows: using the maximum score for the genes within the variant region, GO $(1e-03, 6, 395, 0.126, selu)$, MP $(1e-03, 6, 214, 0.165, sigmoid)$, HP $(1e-04, 6, 391, 0.204, selu)$, CL $(1e-05, 2, 100, 0.2, relu)$, UBERON $(2e-03, 5, 414, 0.316, selu)$, and combined model $(0.002, 6, 70, 0.132, relu)$; and using the average score: GO $(9e-04, 2, 147, 0.278, relu)$, MP $(7e-$

$05, 3, 467, 0.121, relu$), HP ($1e-03, 5, 121, 0.144, relu$), CL ($1e-03, 3, 142, 0.235, relu$), UBERON ($3e-03, 2, 420, 0.104, sigmoid$), and combined model ($3e-04, 3, 439, 0.192, relu$).

For all the models, we train for a maximum of 100 epochs. We use 15% of the training set for validation and stop training if the validation set loss increases for 10 epochs. We use a binary cross-entropy during the training to minimize the loss function. We use a $sigmoid$ activation function for the output layer. We implemented our models using Keras with a TensorFlow backend, and training was performed on a single Nvidia Tesla V100 GPU.

2.5 Benchmarking of structural variant prioritization methods

We compared the performance of DeepSVP to two related methods that can rank structural variants, StrVCTVRE (Sharo *et al.*, 2020), and AnnotSV version 2.3 (Geoffroy *et al.*, 2018). StrVCTVRE is a structural variant impact predictor that captures a set of genomic features for structural variants relating to the conservation, gene importance, coding region, expression, and exon structure, trained using a random forest classifier. AnnotSV provides a classification for each SV based on recommendations for the interpretation of copy number variants (Riggs *et al.*, 2019) and classifies variants into pathogenic, likely pathogenic, of uncertain significance, likely benign, and benign. AnnotSV can use the phenotype-based method Exomiser (Smedley *et al.*, 2015) to determine whether phenotypes are consistent with previously reported cases, and incorporate the phenotype-based score in the variant classification process. We rank variants based on the class assigned by AnnotSV, descending from pathogenic to benign. To compare the prediction performance of our models and other tool, we use the area under the Receiver Operating Characteristics (ROC) curve (AUC), F1-score, the Area Under the Precision–Recall Curve (AUCPR), and the Diagnostic Odds Ratio (DOR) (Glas *et al.*, 2003).

As a benchmark set, we use the disease-associated variants added to the dbVar database (Griffith and Griffith, 2004) between 8 February 2020 and 2 July 2020. Our training dataset and that of StrVCTVRE and AnnotSV are limited to the set of variants that have been added to dbVar prior to that date. 1,503 disease-associated variants were added between 8 February 2020 and 2 July 2020, covering 579 distinct diseases, 175 of which are not linked with any disease in our training set. The evaluation using entirely unseen disease-associated variants helps us to estimate how well DeepSVP can prioritize novel variants under more realistic conditions.

We created synthetic patient samples by inserting a single causative variant into a whole genome sequence from the 1,000 Genomes Project for structural variant (1000 Genomes Project Consortium *et al.*, 2012). The set of structural variants in 1,000 Genomes contains a total of 68,697 variants for 2,504 individuals from 26 populations. Using the 1000 Genomes frequencies for all populations, we exclude all variants with Minor Allele Frequency (MAF) of more than 1% which results in 2,391 variants remaining. Then we link the phenotypes associated with (OMIM) that are associated with the causative variant in dbVar (using `phenotype_annotation.tab`) to the synthetic genome. We consider the combination of the synthetic genome and HPO phenotypes as a synthetic patient sample. We repeat this for all 1,503 causative variants.

2.6 Whole genome sequencing and structural variant calling

We collected blood samples for a Saudi family consisting of five individuals, two unaffected parents, two affected children and one unaffected child. We performed whole genome shotgun sequencing on all individuals. We prepared 150-bp paired-end libraries using the TruSeq Nano DNA Sample Preparation kit (Illumina, USA). Sequencing was performed using an Illumina HiSeq 4000 at the Bioscience core laboratory, KAUST, with approximately 30X coverage.

Following sequencing, we aligned reads to human genome build hg38 using the BWA MEM algorithm (Li, 2013) and following the GATK standard workflows. We trim adapters using Trimmomatic (version 0.38), use BWA (version 0.7.17) for alignment, and samtools (version 1.8) (Li *et al.*, 2009) to remove duplicates and sort bam files. We use Manta (version 1.6) (Chen *et al.*, 2016) to call structural variants. In total, we identified 8,723, 9,003, 9,608, 7,631, and 8,367 variants for the mother, father, first affected, second affected, and unaffected, respectively. We assume an autosomal recessive mode of inheritance or *de novo* variants, and use the pedigree to filter variants. We further filter common variants (minor allele frequency greater than 0.01) using gnomAD (version 2.1.1) (Collins *et al.*, 2020) and the structural variants from the 1000 genomes project (1000 Genomes Project Consortium *et al.*, 2012). After filtering by pedigree, 148 structural variants remain; removing common variants reduced the number of variants to 47. We use the DeepSVP combined model with *maximum* as aggregation operation for the genes within the variant to prioritize disease-associated structural variants.

2.7 Ethical approvals

This study was approved by the Institutional Research Board of the King Abdullah International Medical Research Center (RC 16/113 and RC16/211/R2), and the Institutional Bioethics Committee at King Abdullah University of Science and Technology (17IBEC08_Gojobori). All patients have been consented to be enrolled in this study, a written consent form was obtained from all subjects or their parents or legal guardians in the case of minors who are aged 16 years old or younger.

3 Results

3.1 Predicting disease-associated structural variants

We developed DeepSVP as a method to identify phenotype-associated structural variants (deletions and duplications) for patients based on personal genomic data and the phenotypes observed in a patient. Our method uses the phenotypes arising from a loss of function in mouse, phenotypes associated with human genes, the anatomical site of gene expression, gene functions, and celltypes in which genes are expressed, as background knowledge, and links these to the abnormal phenotypes observed in the individual in which the structural was detected. To make the predictions based on these different features types, we embed them into a shared representation space using a feature learning method applied to ontologies (Chen *et al.*, 2020). Then we combine the resulting embeddings with sequence-derived features that can be used to predict the pathogenicity of a variant, and use a neural network model to predict whether a variant is associated with patient phenotypes. The workflow of DeepSVP is illustrated in Figure 1.

The aim of our model is not only to detect potentially pathogenic structural variants, but identify the variants that are causative of a set of phenotypes observed in a patient. We consider a variant as cause of a set of phenotypes when it is both pathogenic (i.e., disrupts the normal functioning of one or more genes) and contributes directly to the development of the phenotypes. This approach is motivated by the observation that even healthy individuals may have pathogenic or potentially pathogenic variants that do not result in abnormal phenotypes. Therefore, detecting pathogenicity of a variant alone is typically not sufficient to establish causality (MacArthur *et al.*, 2014). Our workflow uses as input a set of variants and a set of phenotypes observed in an individual; for each variant–phenotypes pair, DeepSVP predicts whether the variant is likely to be causative of the phenotypes, and outputs a ranked list of results.

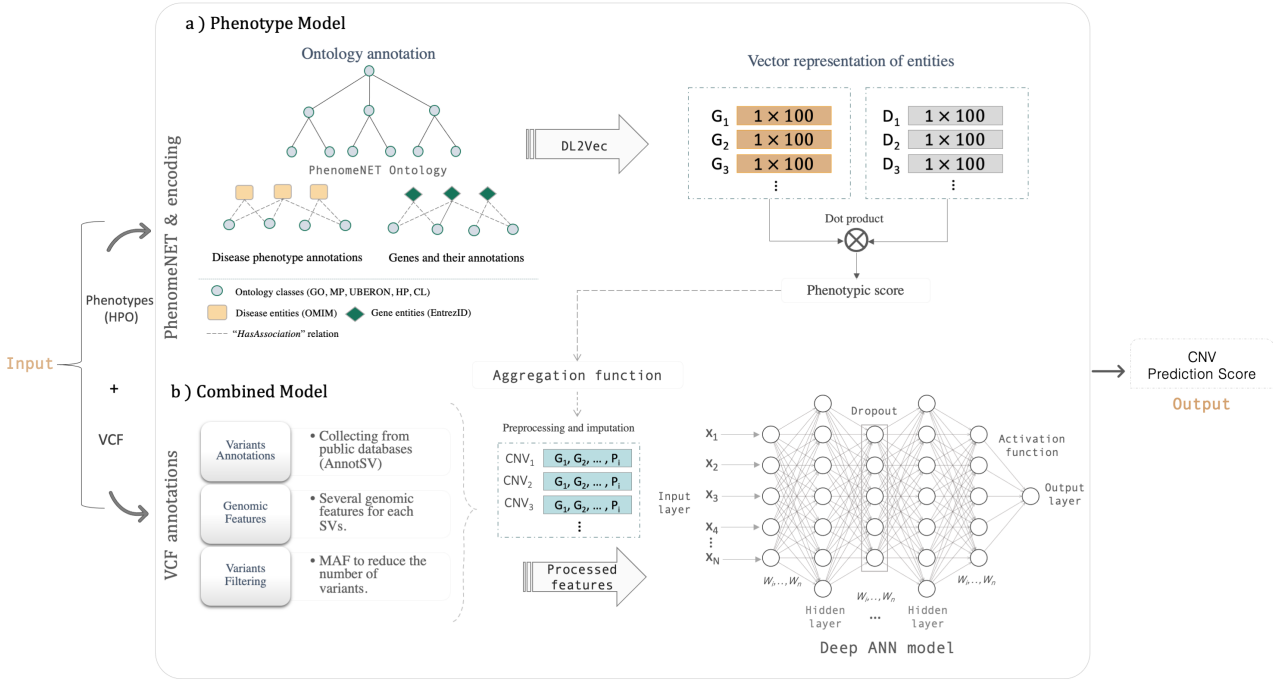

Fig. 1: Overview over DeepSVP model. (**a**) The DL2Vec workflow takes a set of phenotypes as input and predicts whether a gene is likely associated with these phenotypes using several types of background knowledge. (**b**) The combined model uses the prediction score of the DL2Vec model combined with genomic features derived from variants in a VCF file and outputs a prediction score for each variant in the VCF file that determines how likely the variant is causative of the phenotypes observed in the patient.

We used 85% of the diseases in our dataset, together with all their associated structural variants, for training and 15% for testing. We first evaluate DeepSVP's performance in finding disease-causing structural variants on these 15%. Supplementary Table 2 and Supplementary Figure 2 show a summary of DeepSVP's performance. We find that the our model can separate positive from negative cases with a ROCAUC ranging from 0.8913 (using only celltype of expression as background knowledge) to 0.9534 (combining all background knowledge). However, the evaluation on a testing dataset that resembles the training data will not be indicative of the performance of the model in a realistic setting where the aim is to identify a single disease-associated variant among potentially hundreds or thousands of candidates within a genome.

As a more realistic evaluation of our model, we generate synthetic patient data in which we combine the variants from the genome sequences in the 1,000 Genomes project, insert a single disease-causing pathogenic variant, and associate this synthetic genome with the phenotypes of the variant. We then apply our model to all structural variants in this synthetic patient, rank the resulting variants based on DeepSVP's prediction score, and evaluate the results. For this evaluation, we select an independent dataset of disease-associated structural variants, i.e., the set of variant–disease pairs added to dbVar (Griffith and Griffith, 2004) after we obtained training data for our model. Our evaluation set contains 1,503 variants associated with 579 distinct diseases in OMIM and overlapping with 1,926 unique genes. There are 175 diseases (associated with 640 variants) that were not present in our training data. We create synthetic patient samples for all variant–phenotype pairs in this evaluation set. We also compare the results of our model with another method for identifying disease-associated variants, StrVCTVRE (Sharo *et al.*, 2020), and AnnotSV (Geoffroy *et al.*, 2018). Table 1 shows the amount of disease-associated variants we identify at different ranks. We evaluate the recall separately for variants that are associated with a disease already present in our training data and variants that were not.

The performance using different DeepSVP models varies depending on the availability of the features as well as the gene annotations that are available for each type of feature. The benchmark dataset (i.e., the list of variants we inserted) is linked with 1,920 unique genes, with the largest variant containing 129 genes and the smallest variant containing one gene. There are $635, 963, 1214, 1360, 493$ missing annotations from these genes for features represented using GO, MP, HP, CL, and UBERON, respectively. However, the models are still able to predict causative variants using the annotations of the remaining genes ($1285, 957, 706, 560, 1427$ for GO, MP, HP, CL, and UBERON) which have an annotation and a corresponding "representation". The evaluation demonstrates that DeepSVP can identify the variants in novel (i.e., unseen during training) disease-associated variants with one, or more than one, gene and still improve recall at rank one over some established methods. We also find that DeepSVP significantly improves prioritization of disease-associated variants over StrVCTVRE ($p < 3.9 \times 10^{-13}$, Mann-Whitney U test), a method that uses similar features as DeepSVP to determine pathogenicity of variants but does not rely on information about phenotypic or functional consequences. We further evaluated the rank provided by AnnotSV which classifies variants into five classes (pathogenic, likely pathogenic, unknown significance, likely benign, benign); while AnnotSV has the highest recall at rank 1 (i.e., variants predicted as "pathogenic"), it also predicts a large number of other variants at the same rank (i.e., as pathogenic variants), leading to a lower precision compared to DeepSVP or StrVCTVRE which rank variants without ties.

While we are able to identify disease-associated structural variants using phenotype information, the phenotypes reported with a patient-derived sample will not always be complete. To evaluate the effect of

		Synthetic dataset				Synthetic dataset (novel diseases)			
Aggregation method for genes within variant	**DeepSVP Models**	**Recall@1**	**Recall@10**	**Recall@30**	**Recall@50**	**Recall@1**	**Recall@10**	**Recall@30**	**Recall@50**
Maximum score	GO	27 (0.018)	538 (0.358)	821 (0.5462)	953 (0.6341)	20 (0.0312)	295 (0.4609)	402 (0.6281)	438 (0.6844)
	MP	65 (0.0432)	986 (0.656)	1144 (0.7611)	1207 (0.8031)	13 (0.0203)	451 (0.7047)	518 (0.8094)	535 (0.8359)
	HP	335 (0.2229)	1093 (0.7272)	**1369 (0.9108)**	1387 (0.9228)	133 (0.2078)	485 (0.7578)	609 (0.9516)	612 (0.9563)
	CL	23 (0.0153)	516 (0.3433)	868 (0.5775)	982 (0.6534)	23 (0.0359)	234 (0.3656)	299 (0.4672)	344 (0.5375)
	UBERON	73 (0.0486)	531 (0.3533)	629 (0.4185)	719 (0.4784)	39 (0.0609)	318 (0.4969)	359 (0.5609)	377 (0.5891)
	GO+MP+HP+CL+UBERON	172 (0.1144)	867 (0.5768)	953 (0.6341)	1063 (0.7073)	37 (0.0578)	404 (0.6312)	437 (0.6828)	504 (0.7875)
Average score	GO	236 (0.1570)	992 (0.6600)	1333 (0.8869)	1343 (0.8935)	72 (0.1125)	442 (0.6906)	600 (0.9375)	605 (0.9453)
	MP	28 (0.0186)	916 (0.6094)	1332 (0.8862)	1339 (0.8909)	18 (0.0281)	407 (0.6359)	598 (0.9344)	602 (0.9406)
	HP	76 (0.0506)	671 (0.4464)	1340 (0.8916)	1352 (0.8995)	30 (0.0469)	381 (0.5953)	606 (0.9469)	610 (0.9531)
	CL	287 (0.1910)	984 (0.6547)	1359 (0.9042)	1388 (0.9235)	116 (0.1812)	449 (0.7016)	609 (0.9516)	613 (0.9578)
	UBERON	488 (0.3247)	**1128 (0.7505)**	1358 (0.9035)	1370 (0.9115)	144 (0.225)	**509 (0.7953)**	**610 (0.9531)**	612 (0.9563)
	GO+MP+HP+CL+UBERON	65 (0.0432)	726 (0.4830)	897 (0.5968)	1044 (0.6946)	30 (0.0469)	422 (0.6594)	555 (0.8672)	590 (0.9219)
SV pathogenicity prediction/ranking	StrVCTVRE	38 (0.0252)	620 (0.4125)	1022 (0.6799)	1343 (0.8935)	9 (0.0140)	162 (0.2531)	373 (0.5828)	552 (0.8625)
	AnnotSV	**930 (0.6188) (pr: 0.01561)**	930 (0.6188) (pr: 0.01561)	930 (0.6188) (pr: 0.01561)	**1493 (0.9933) (pr: 0.0026)**	**248 (0.3875) (pr: 0.0098)**	248 (0.3875) (pr: 0.0098)	248 (0.3875) (pr: 0.0098)	**636 (0.9938) (pr: 0.0026)**

Table 1. Summary of the evaluation for predicting causative variants in the benchmark dataset of dbVar, time-based split, for 1,503 newly added variants along with the evaluation for 640 newly added variants associated with 175 new diseases which were not present in our training dataset. The evaluation inserts one disease-associated structural variant in a whole-genome and reports the rank at which the inserted variant was recovered. We report the absolute number of variants recovered at each rank; the numbers in parentheses are the recall at the rank; AnnotSV contains variants that are predicted at the same rank and we report the precision (indicated by "pr:") in addition to the recall at ranks; other methods do not contain ties, and precision corresponds directly to the rank (minus correctly predicted variants). Highest recall is indicated in bold.

different phenotype associations, we further evaluated the performance of DeepSVP when only partial phenotype data is available. We repeat our experiment using synthetic patient samples while randomly removing between 10% and 50% of the phenotypes associated with the sample (Supplementary Tables 3, 4, 5, and 6). We find that even when reducing the number of associated phenotypes to 50%, the performance of our model remains comparable; however, when removing phenotype information entirely using the combined model, the predictive performance drops compared to a model that includes information about phenotypes, which results in $27(1.8\%)$ of causative variants ranked first and $575(38,26\%)$ in the top 10.

3.2 Identification of disease-associated variants in consanguineous families

We applied DeepSVP to investigate whole genome sequencing data in nine consanguineous families from Saudi Arabia (Alfares *et al.*, 2020) where the clinical presentation is suggestive of genetics underlying etiology with two or more affected individuals, and all previous genetic analyses based on clinical and whole exome analysis were negative. First, we investigated single nucleotide variants and small insertions and deletions as possible explanation of the underlying genetics etiology. However, we were unable to identify a possible disease-associated variant in all nine families using this approach (Alfares *et al.*, 2020).

We applied DeepSVP to rank structural variants for possible explanation of the phenotype. In one family, the affected individuals showed hypotonia (`HP:0001290`), developmental delay (`HP:0001263`), infantile spasms (`HP:0012469`), strabismus (`HP:0000486`), and seizures (`HP:0001250`); after removing all common variations, we ended up with 47 structural variants requiring further investigation. DeepSVP ranked one duplication in chromosome `2q24.3`(`NC_000002.12:g.164062341_166264282dup`) as top rank using the combined model based on the maximum score (see Supplementary Table 7 for other DeepSVP models); this duplication contains 16 genes (*COBLL1*, *CSRNP3*, *GALNT3*, *GRB14*, *LOC100506124*, *LOC101929633*, *LOC102724058*, *SCN1A*, *SCN1A-AS1*, *SCN2A*, *SCN3A*, *SCN9A*, *SLC38A11*, *SNORA70F*, *TTC21B*, *TTC21B-AS1*). Duplication of 2q24.3, including the cluster of voltage-gated sodium channel genes, is linked with hypotonia, seizures, and neonatal epilepsy in several unrelated cases (Simonetti *et al.*, 2012; Okumura *et al.*, 2011; Firth *et al.*, 2009). We further confirmed the variant with length 2201941 using Array Comparative Genomic Hybridization in a clinical laboratory.

DeepSVP outputs the gene based on which phenotypic similarity was established (using the maximum phenotype score), and among the 16 genes, the *SCN1A* gene has the maximum phenotype score using MP model. The heterozygote loss of function of *SCN1A* in mouse is a model of Dravet syndrome (Kim *et al.*, 2018; Miller *et al.*, 2013) and resembles many of the phenotypes observed in the family we analyzed. Our results shows that DeepSVP can not only identify disease-associated structural variants but further yield interpretable results that can provide actionable clinical information. We compared the results with predictions of AnnotSV and StrVCRTRE. AnnotSV classified 11 variants out of 47 as pathogenic, including the variant we identified. StrVCRTRE scored only 6 variants out of 47 (the remaining variants are either less than $50bp$ in length or not exonic, and thereby out of the scope of StrVCRTRE), and the variants we identified ranked 4 out of 6.

4 Discussion

4.1 Related work

A seminal study investigated the application of phenotype-similarity in CNVs (Doelken *et al.*, 2012) and the results let to the PhenoGramViz method and tool (Köhler *et al.*, 2014). PhenoGramViz, as well as AnnotSV (Geoffroy *et al.*, 2018), use phenotypes to rank and prioritize structural variants using phenotype information. To overcome the limitation of missing information about phenotypes, PhenoGramViz and AnnotSV rely on mouse phenotypes. While mouse phenotypes increase the coverage of genes with phenotype associations, there are nevertheless a large number of genes for which no phenotype associations are available. DeepSVP overcomes the limitation of missing phenotypes by incorporating information related to genes through ontologies, mainly the functions of gene products, gene expression in individual celltypes, and anatomical sites of expression and systematically relating them to their phenotypic consequences through ontologies.

PhenoGramViz is not a method that directly prioritizes structural variants but relies on visualizing ranking results and exploration by users. While this is useful in targeted studies, DeepSVP can be applied as a component of computational workflows while still enabling interpretation of results. AnnotSV provides a classification rank for each SV using five classes based on their overlap with known variants from different data sources, and aims to implement clinical classification guidelines for variants. DeepSVP, on the other hand, provides pathogenicity prediction for each variant rather than categorize them and includes phenotype prediction models not only to identify relatedness to known phenotypes but also to predict new ones; it may therefore be more suitable for generating hypotheses about phenotype-associations of structural variants that do not overlap with known disease genes. StrVCRTRE (Sharo *et al.*, 2020) is a method that also directly predicts pathogenicity of structural variants and uses similar features related to the gene importance, coding sequence, and expression, which allows us to compare directly. A key difference between DeepSVP and StrVCRTRE is DeepSVP's use of phenotype information while StrVCRTRE does not rely on phenotype information which improves prediction results significantly. Furthermore, StrVCRTRE ranks only the exonic variants, while DeepSVP ranks both exonic and intronic based on the availability of the genomics features.

4.2 Machine learning with semantic background knowledge: from transductive to inductive

DeepSVP relies on machine learning for predicting pathogenic and phenotype-associated with the patient. For this purpose, it relies on advances in machine learning with ontologies that incorporate the background knowledge contained in ontologies in the form of axioms and annotations to ontology classes (Kulmanov *et al.*, 2020). Many such approaches convert ontologies into a graph-based form based on syntactic patterns within the ontology axioms and then apply a graph embedding on the resulting graph (Kulmanov *et al.*, 2020). In DeepSVP, we use DL2Vec which includes a large variety of ontology axioms and can significantly improve the phenotype–based prediction of disease genes (Chen *et al.*, 2020).

While these ontology-based methods rank genes, DeepSVP directly ranks structural variants based on the genomic and phenotypic features collected from public databases, and the phenotypes observed in a patient. We precomputed the embeddings for genes based on different features (function, phenotype, and expression in celltypes and anatomical parts). Furthermore, we extended the ontology-based machine learning methods to an inductive setting where we can predict associations between genes and patients that are defined by their phenotypes which are not known at the time of training DeepSVP. We also applied a rank-based normalization, similar to the method applied by PhenoRank (Cornish *et al.*, 2018), and use the resulting score instead of prediction scores of the neural network model; this transformation is useful when predicting relations where one argument remains fixed as it projects prediction scores into the same distribution.

While we implemented a two-step approach in which we first predict associations between genes and patient phenotpyes, and second the pathogenicity and phenotypic relatedness of the variant to the patient phenotypes, it may also be possible to design a model that is trained in an end-to-end fashion in the future. The challenge is the potentially open-ended number of genes to consider.

4.3 Clinical application and utility

We evaluated the performance of DeepSVP on a series of real genomes from Saudi individuals where the clinical presentation is suggestive of genetic diseases to assess how well we could recover potentially pathological variants in genes already associated with the disease. We used the whole genome sequencing data from all family members, and we apply family filtering according to the suitable inheritance pattern. We applied DeepSVP to rank structural variants for a possible explanation of the phenotype. In one family, our model was able to find the causative variants associated with the patient phenotypes using the combined prediction model that integrates all the phenotypes information, and also to highlight a candidate gene underlying the main phenotypic manifestations.

We implemented two models for aggregating phenotypic relatedness between genes and patient phenotypes, one using the maximum and another using the average of scores of all genes. These correspond to two different mechanisms through which a structural variant elicits abnormal phenotypes: the maximum model is applicable when a single gene within the variant is (primarily) causative for the phenotypes, whereas the average model is applicable in the oligogenic case when multiple genes affected by a structural variant are causative and may contribute different pathologies.

We make DeepSVP freely available to use as a free software command-line tool, including all the steps to train the model. DeepSVP uses as input an annotated Variant Call Format (VCF) file of an individual and clinical phenotypes encoded using the Human Phenotype Ontology (HPO). It can be used as a part of interpretation workflows in a clinical setting, or incorporated in interactive variant exploration methods.

Acknowledgements

This work was supported by funding from King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3790-01-01, FCC/1/1976-08-01, and FCC/1/1976-08-08. We acknowledge support from the KAUST Supercomputing Laboratory.

References

1000 Genomes Project Consortium *et al.* (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**(7422), 56.

Alfares, A. *et al.* (2020). What is the right sequencing approach? solo VS extended family analysis in consanguineous populations. *BMC Medical Genomics*, **13**(1).

Amberger, J. *et al.* (2011). A new face and new challenges for online mendelian inheritance in man (omim®). *Human mutation*, **32**(5), 564–567.

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, **75**(4), 761–771.

Bult, C. J. *et al.* (2019). Mouse genome database (mgd) 2019. *Nucleic acids research*, **47**(D1), D801–D806.

Chen, J. *et al.* (2020). Predicting candidate genes from phenotypes, functions and anatomical site of expression. *Bioinformatics*. btaa879.

Chen, X. *et al.* (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, **32**(8), 1220–1222.

Collins, R. L. *et al.* (2020). A structural variation reference for medical and population genetics. *Nature*, **581**(7809), 444–451.

Cornish, A. J. *et al.* (2018). PhenoRank: reducing study bias in gene prioritization through simulation. *Bioinformatics*, **34**(12), 2087–2095.

Diehl, A. D. *et al.* (2016). The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of biomedical semantics*, **7**(1), 1–10.

Doelken, S. C. *et al.* (2012). Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish. *Disease Models & Mechanisms*, **6**(2), 358–372.

Eichler, E. E. (2019). Genetic variation, comparative genomics, and the diagnosis of disease. *New England Journal of Medicine*, **381**(1), 64–74.

Eilbeck, K. *et al.* (2017). Settling the score: variant prioritization and mendelian disease. *Nature Reviews Genetics*, **18**(10), 599–612.

Firth, H. V. and Wright, C. F. (2011). The deciphering developmental disorders (DDD) study. *Developmental medicine and child neurology*, **53**(8), 702.

Firth, H. V. *et al.* (2009). Decipher: database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics*, **84**(4), 524–533.

Ganel, L. *et al.* (2017). Svscore: an impact prediction tool for structural variation. *Bioinformatics*, **33**(7), 1083–1085.

Gene Ontology Consortium (2019). The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, **47**(D1), D330–D338.

Geoffroy, V. *et al.* (2018). AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*, **34**(20), 3572–3574.

Glas, A. S. *et al.* (2003). The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology*, **56**(11), 1129–1135.

Griffith, M. and Griffith, O. L. (2004). dbVar (Database of Genomic Structural Variation). *Dictionary of Bioinformatics and Computational Biology*.

GTEx Consortium (2015). The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, **348**(6235), 648–660.

Hehir-Kwa, J. Y. *et al.* (2010). Accurate distinction of pathogenic from benign CNVs in mental retardation. *PLoS computational biology*, **6**(4), e1000752.

Huntley, R. P. *et al.* (2014). The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Research*, **43**(D1), D1057–D1063.

Karczewski, K. J. *et al.* (2017). The exac browser: displaying reference data information from over 60 000 exomes. *Nucleic acids research*, **45**(D1), D840–D845.

Kidd, J. M. *et al.* (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**(7191), 56–64.

Kim, Y. *et al.* (2018). Severe peri-ictal respiratory dysfunction is common in dravet syndrome. *Journal of Clinical Investigation*, **128**(3), 1141–1153.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Köhler, S. *et al.* (2014). Clinical interpretation of CNVs with cross-species phenotype data. *Journal of medical genetics*, **51**(11), 766–772.

Köhler, S. *et al.* (2019). Expansion of the human phenotype ontology (hpo) knowledge base and resources. *Nucleic acids research*, **47**(D1), D1018–D1027.

Kulmanov, M. and Hoehndorf, R. (2020). DeepPheno: Predicting single gene loss-of-function phenotypes using an ontology-aware hierarchical classifier. *PLOS Computational Biology*, **16**(11), e1008453.

Kulmanov, M. *et al.* (2020). Semantic similarity and machine learning with ontologies. *Briefings in Bioinformatics*. bbaa199.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.

Li, H. *et al.* (2009). The sequence alignment/map format and samtools. *Bioinformatics*, **25**(16), 2078–2079.

Maas, A. L. *et al.* (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3.

MacArthur, D. *et al.* (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, **508**(7497), 469–476.

MacDonald, J. R. *et al.* (2014). The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic acids research*, **42**(D1), D986–D992.

Maglott, D. *et al.* (2010). Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, **39**(suppl_1), D52–D57.

Mikolov, T. *et al.* (2013). Efficient estimation of word representations in vector space. *CoRR*, **abs/1301.3781**.

Miller, A. R. *et al.* (2013). Mapping genetic modifiers of survival in a mouse model of dravet syndrome. *Genes, Brain and Behavior*, **13**(2), 163–172.

Mungall, C. J. *et al.* (2012). Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, **13**(1), R5.

Okumura, A. *et al.* (2011). Refractory neonatal epilepsy with a de novo duplication of chromosome 2q24.2q24.3. *Epilepsia*, **52**(7), e66–e69.

O'Leary, N. A. *et al.* (2016). Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, **44**(D1), D733–D745.

Papatheodorou, I. *et al.* (2020). Expression atlas update: from tissues to single cells. *Nucleic Acids Research*, **48**(D1), D77–D83.

Pinto, D. *et al.* (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**(7304), 368.

Rehm, H. L. *et al.* (2015). Clingen—the clinical genome resource. *New England Journal of Medicine*, **372**(23), 2235–2242.

Riggs, E. R. *et al.* (2019). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the american college of medical genetics and genomics (ACMG) and the clinical genome resource (ClinGen). *Genetics in Medicine*, **22**(2), 245–257.

Rodríguez-García, M. Á. *et al.* (2017). Integrating phenotype ontologies with phenomenet. *Journal of biomedical semantics*, **8**(1), 58.

Rossin, E. J. *et al.* (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS genetics*, **7**(1), e1001273.

Sanchis-Juan, A. *et al.* (2018). Complex structural variants in mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Medicine*, **10**(1).

Sharo, A. G. *et al.* (2020). StrVCTVRE: A supervised learning method to predict the pathogenicity of human structural variants. *BioRxiv*.

Shefchek, K. A. *et al.* (2020). The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic acids research*, **48**(D1), D704–D715.

Simonetti, B. G. *et al.* (2012). Duplication of the sodium channel gene cluster on 2q24 in children with early onset epilepsy. *Epilepsia*, **53**(12), 2128–2134.

Smaili, F. Z. *et al.* (2019). Opa2vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics*, **35**(12), 2133–2140.

Smedley, D. *et al.* (2013). Phenodigm: analyzing curated annotations to associate animal models with human diseases. *Database*, **2013**.

Smedley, D. *et al.* (2014). Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*, **30**(22), 3215–3222.

Smedley, D. *et al.* (2015). Next-generation diagnostics and disease-gene discovery with the exomiser. *Nature protocols*, **10**(12), 2004–2015.

Sudmant, P. H. *et al.* (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**(7571), 75–81.

Tabula Muris Consortium *et al.* (2018). Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, **562**(7727), 367.

UniProt Consortium (2019). Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, **47**(D1), D506–D515.

Zhou, N. *et al.* (2019). The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, **20**(1), 1–23.

DeepSVP: Integration of Genomics and Phenotypes for Structural Variant Prioritization using Deep Learning (Supplementary materials)

Azza Althagafi[1,2], Lamia Alsubaie[3], Nagarajan Kathiresan[4], Katsuhiko Mineta[1], Taghrid Aloraini[3], Fuad Almutairi[7,8], Majid Alfadhel[5,7,8], Takashi Gojobori[9], Ahmad Alfares[3,5,6], and Robert Hoehndorf[1,*]

[1] Computer, Electrical and Mathematical Sciences & Engineering Division, Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology, 4700 King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

[2] Computer Science Department, College of Computers and Information Technology, Taif University, Taif 26571, Saudi Arabia.

[3] Department of Pathology and Laboratory Medicine, King Abdulaziz Medical City, Riyadh, Saudi Arabia.

[4] Supercomputing Core Lab, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia.

[5] King Abdullah International Medical Research Center, Riyadh, Saudi Arabia.

[6] Department of Pediatrics, College of Medicine, Qassim University, Qassim, Saudi Arabia.

[7] Division of Genetics, Department of Pediatrics, King Abdulaziz Medical City, Riyadh, Saudi Arabia.

[8] King Saud bin Abdulaziz University for Health Sciences, King Abdulaziz Medical City, Riyadh, Saudi Arabia.

[9] Biological and Environmental Science and Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia.

[*] robert.hoehndorf@kaust.edu.sa

	Feature	Description
Structural Variant (SV)	SV length	Length of SV.
	SV type	Type of the SV (DEL, DUP) Categorical features.
	GCcontent_left*	Breakpoints annotations, GC content around the left SV breakpoint (+/- 100bp).
	GCcontent_right*	Breakpoints annotations, GC content around the right SV breakpoint (+/- 100bp).
	Number of genes*	Number of genes within the SVs.
	Number of promoters*	Number of promoters within the SVs.
	HI_CGscore	HaploInsufficiency Score (categorical features).
	TriS_CGscore	TriploSensitivity Score (categorical feature).
Genes-based annotations	DL2vec_Score*	Predict associations between genes and sets of phenotypes different ontolgies (GO, HP, CL, MP, and UBERON).
	CDS length*	Length of the CoDing Sequence (CDS) (bp) overlapping with the SV.
	tx length*	Length of transcript (bp) overlapping with the SV.
	synZ_ExAC*	Gene intolerance to synonymous variation.
	misZ_AxAC*	Gene intolerance to missense variation.
	pLI_ExAC*	The probability that a gene is intolerant to a loss of function variation.
	dupZ_ExAC*	Gene duplication intolerance.
	delZ_ExAC*	Gene deletion intolerance.
	cnvZ_ExAC*	Gne CNV intolerance.
	HI_DDDpercent*	Haploinsufficiency ranks, where in a single functional copy of a gene is insufficient to maintain normal function.

Table 1: Annotation features for model training and prediction. An asterisk (*) indicates that a boolean indicator variable was created in order to handle undefined values for that feature.

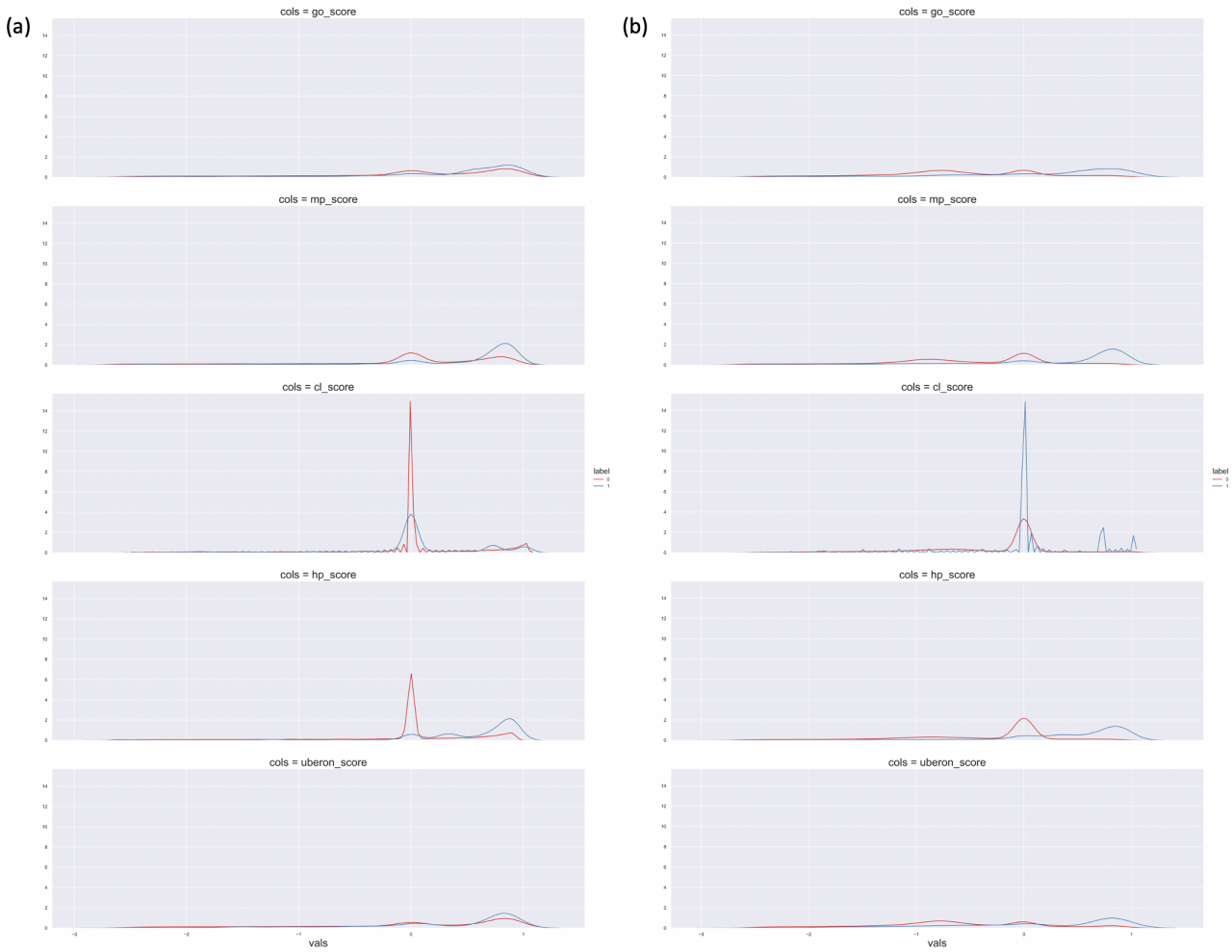

Figure 1: Distribution of the phenotype features, **(a)** using the maximum features scores, and **(b)** using the average score.

Aggregation method for genes within CNV	DeepSVP Models	ROCAUC	F1-Score	PRAUC	DOR
Maximum score	GO	0.9390 ± 0.0068	0.6681 ± 0.0127	0.8126 ± 0.0109	56.1942 ± 0.1290
	MP	0.9293 ± 0.0078	0.6900 ± 0.0134	0.7616 ± 0.0126	41.9207 ± 0.1075
	HP	0.9361 ± 0.0073	0.7009 ± 0.0129	0.7980 ± 0.0116	50.0758 ± 0.1132
	CL	0.9157 ± 0.0083	0.6394 ± 0.0134	0.7581 ± 0.0124	31.6137 ± 0.1067
	UBERON	0.9152 ± 0.0084	0.6011 ± 0.0137	0.7257 ± 0.0130	23.8252 ± 0.1027
	GO+MP+HP+CL+UBERON	**0.9534 ± 0.0062**	**0.7397 ± 0.0125**	**0.8446 ± 0.0105**	**67.9658 ± 0.1177**
Average score	GO	0.9186 ± 0.0081	0.6442 ± 0.0132	0.7189 ± 0.0127	35.4152 ± 0.1108
	MP	0.9336 ± 0.0074	0.6777 ± 0.0132	0.7801 ± 0.0120	41.3731 ± 0.1099
	HP	0.9373 ± 0.0072	0.7077 ± 0.0129	0.8287 ± 0.0110	51.1215 ± 0.1124
	CL	0.8913 ± 0.0093	0.5419 ± 0.0135	0.6656 ± 0.0134	17.3593 ± 0.1020
	UBERON	0.9040 ± 0.0089	0.5698 ± 0.0136	0.7029 ± 0.0132	20.0705 ± 0.1021
	GO+MP+HP+CL+UBERON	**0.9505 ± 0.0062**	**0.7258 ± 0.0123**	**0.8520 ± 0.0101**	**73.3047 ± 0.1262**

Table 2: Summary of the evaluation for predicting causative variants in our testing data set

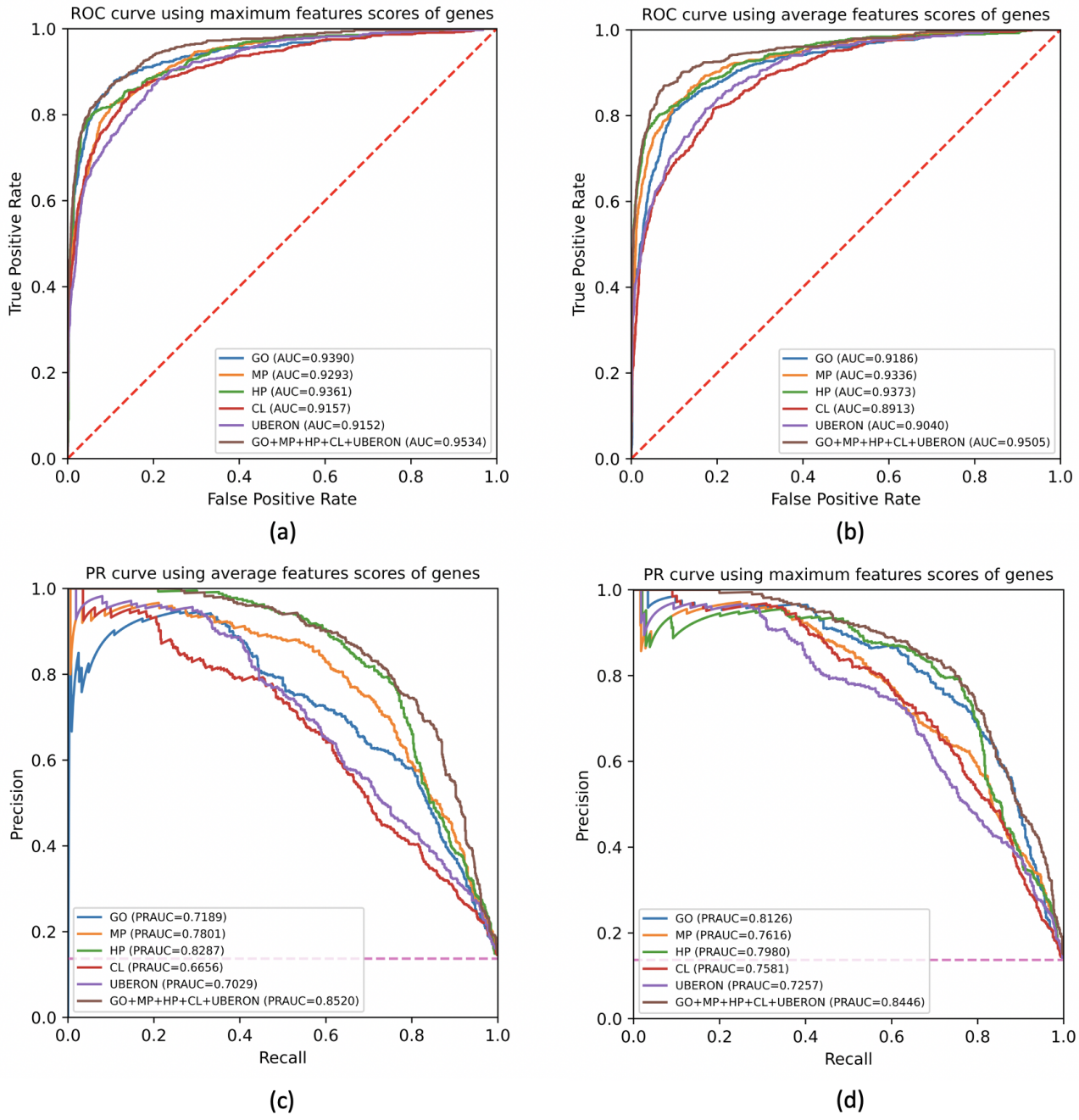

Figure 2: Summary of the ROCAUCs and PRAUCs curves for predicting causative variants using different features and ontologies, and different operation types. **(a)** the ROC curve using the maximum features scores; **(b)** the ROC curve using the average features scores; **(c)** the PR curve using the average features scores; and **(d)** the PR curve using the maximum features scores.

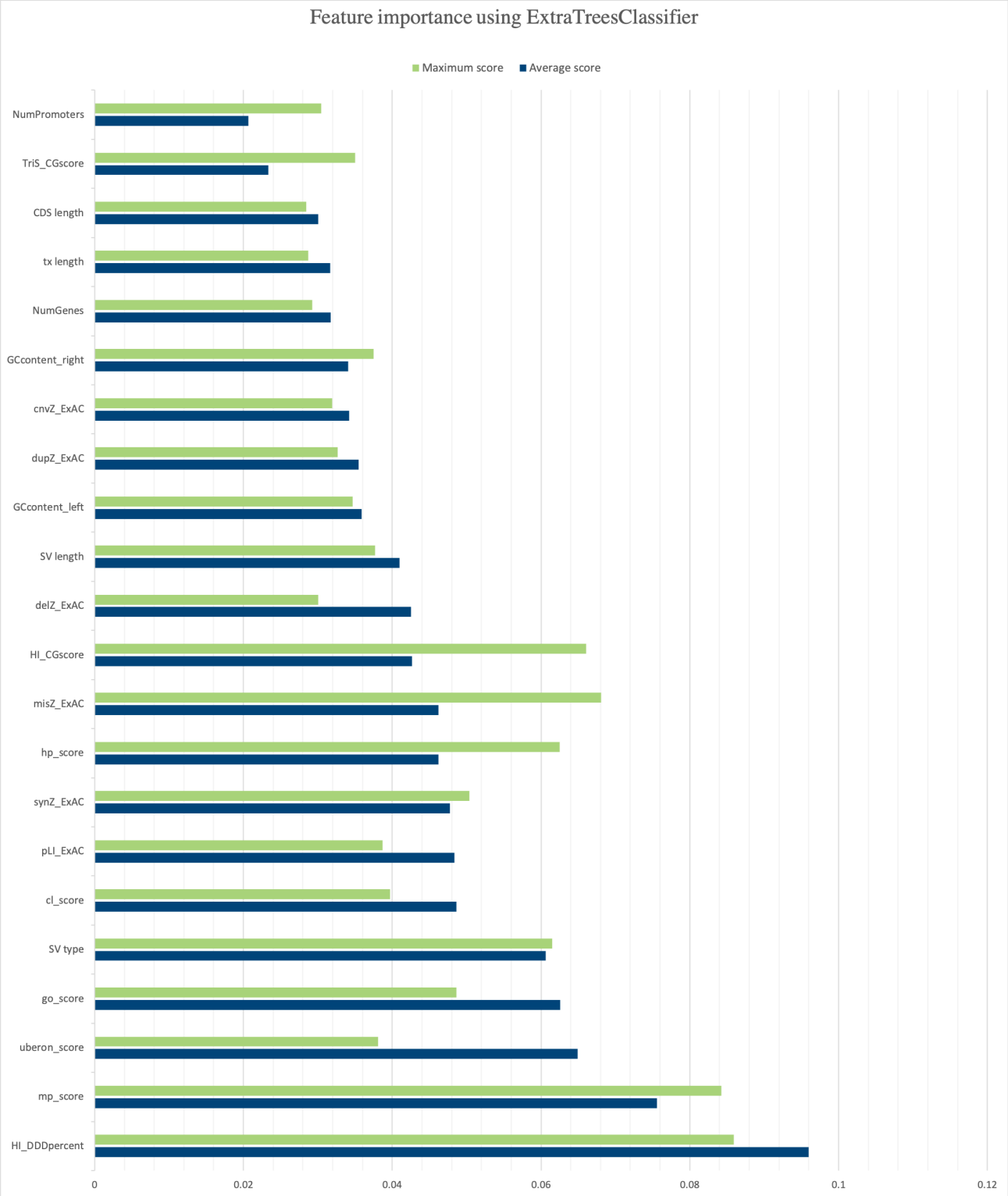

Figure 3: Feature importance using ensemble learning technique Extremely Randomized Trees Classifier (Extra Trees Classifier) that aggregates the results of multiple decision trees and output the ranked features based on the information gain.

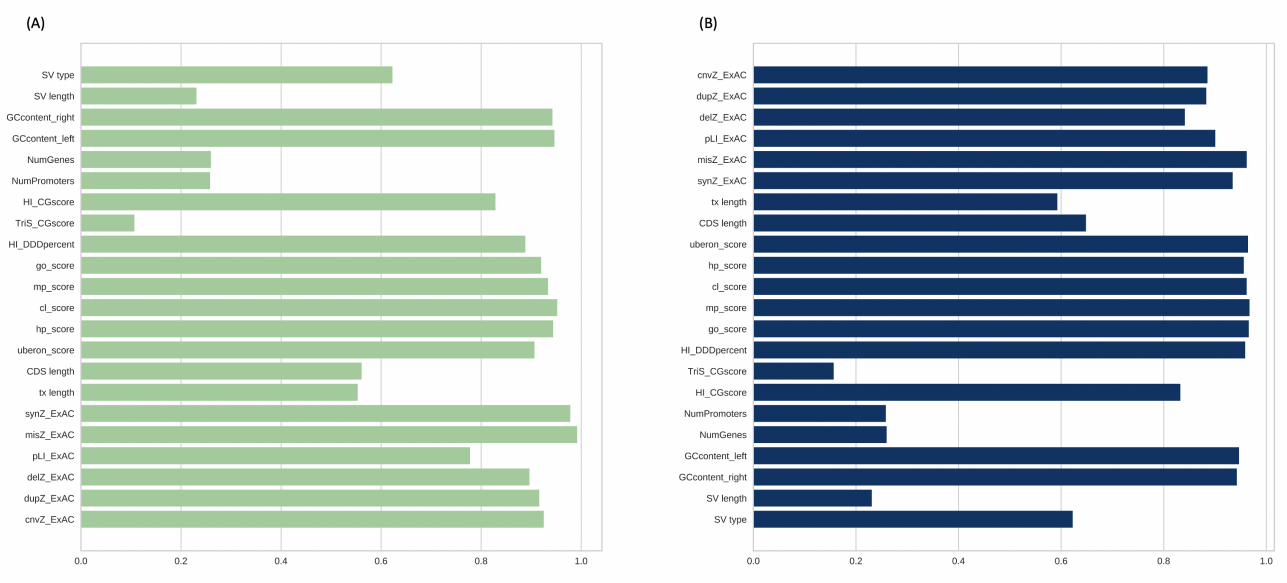

Figure 4: A one-dimensional ranking of features utilizing the Shapiro-Wilk algorithm generated using *Yellowbrick* Python package (version 1.0). The Shapiro algorithm takes into account a single feature at a time to assess the normality of the distribution of instances with respect to the feature. A barplot showing the relative ranks of each feature **(A)** using the maximum features scores, and **(B)** using the average score.

		Synthetic dataset excluding 10% randomly of the phenotypes			
Aggregation method for genes within CNV	**DeepSVP Models**	**Recall@1**	**Recall@10**	**Recall@30**	**Recall@50**
Maximum score	GO	27 (0.0180)	538 (0.3580)	826 (0.5496)	956 (0.6361)
	MP	65 (0.0432)	985 (0.6554)	1141 (0.7591)	1207 (0.8031)
	HP	336 (0.2236)	1093 (0.7272)	**1366 (0.9088)**	1381 (0.9188)
	CL	22 (0.0146)	520 (0.3460)	873 (0.5808)	987 (0.6567)
	UBERON	74 (0.0492)	528 (0.3513)	624 (0.4152)	713 (0.4744)
	GO+MP+HP+CL+UBERON	173 (0.1151)	834 (0.5549)	917 (0.6101)	1039 (0.6913)
Average score	GO	232 (0.1544)	993 (0.6607)	1332 (0.8862)	1343 (0.8935)
	MP	27 (0.0180)	913 (0.6075)	1333 (0.8869)	1341 (0.8922)
	HP	74 (0.0492)	668 (0.4444)	1340 (0.8916)	1350 (0.8982)
	CL	290 (0.1929)	984 (0.6547)	1360 (0.9049)	1390 (0.9248)
	UBERON	488 (0.3247)	**1134 (0.7545)**	1358 (0.9035)	1371 (0.9122)
	GO+MP+HP+CL+UBERON	71 (0.0472)	725 (0.4824)	908 (0.6041)	1060 (0.7053)
SV pathogenicity prediction/ranking	StrVCTVRE	38 (0.0252)	620 (0.4125)	1022 (0.6799)	1343 (0.8935)
	AnnotSV	**930 (0.6188)**	930 (0.6188)	930 (0.6188)	**1493 (0.9933)**

Table 3: Summary of the evaluation for predicting causative variants in the synthetic whole genome benchmark dataset derived from dbVar, using 90% of the phenotypes.

		Synthetic data set excluding 20% randomly of the phenotypes			
Aggregation method for genes within CNV	**PredCNV Models**	Recall@1	Recall@10	Recall@30	Recall@50
Maximum score	GO	27 (0.018)	538 (0.358)	827 (0.5502)	957 (0.6367)
	MP	70 (0.0466)	984 (0.6547)	1146 (0.7625)	1207 (0.8031)
	HP	332 (0.2209)	1091 (0.7259)	**1371 (0.9122)**	1388 (0.9235)
	CL	23 (0.0153)	519 (0.3453)	870 (0.5788)	987 (0.6567)
	UBERON	74 (0.0492)	527 (0.3506)	625 (0.4158)	713 (0.4744)
	GO+MP+HP+CL+UBERON	167 (0.1111)	866 (0.5762)	949 (0.6314)	1062 (0.7066)
Average score	GO	233 (0.155)	993 (0.6607)	1332 (0.8862)	1344 (0.8942)
	MP	27 (0.018)	912 (0.6068)	1333 (0.8869)	1340 (0.8916)
	HP	72 (0.0479)	667 (0.4438)	1342 (0.8929)	1351 (0.8989)
	CL	290 (0.1929)	988 (0.6574)	1360 (0.9049)	1388 (0.9235)
	UBERON	488 (0.3247)	**1133 (0.7538)**	1358 (0.9035)	1372 (0.9128)
	GO+MP+HP+CL+UBERON	71 (0.0472)	725 (0.4824)	908 (0.6041)	1060 (0.7053)
SV pathogenicity prediction/ranking	StrVCTVRE	38 (0.0252)	620 (0.4125)	1022 (0.6799)	1343 (0.8935)
	AnnotSV	**930 (0.6188)**	930 (0.6188)	930 (0.6188)	**1493 (0.9933)**

Table 4: Summary of the evaluation for predicting causative variants in the synthetic whole genome benchmark dataset derived from dbVar, using 80% of the phenotypes.

		Synthetic dataset excluding 30% randomly of the phenotypes			
Aggregation method for genes within CNV	**DeepSVP Models**	**Recall@1**	**Recall@10**	**Recall@30**	**Recall@50**
Maximum score	GO	27 (0.01800)	538 (0.3580)	827 (0.5502)	955 (0.6354)
	MP	67 (0.0446)	983 (0.6540)	1139 (0.7578)	1203 (0.8004)
	HP	334 (0.2222)	1093 (0.7272)	1373 (0.9135)	1384 (0.9208)
	CL	23 (0.0153)	519 (0.3453)	871 (0.5795)	992 (0.6600)
	UBERON	75 (0.0499)	529 (0.3520)	627 (0.4172)	716 (0.4764)
	GO+MP+HP+CL+UBERON	177 (0.1178)	864 (0.5749)	958 (0.6374)	1057 (0.7033)
Average score	GO	234 (0.1557)	990 (0.6587)	1331 (0.8856)	1344 (0.8942)
	MP	28 (0.0186)	912 (0.6068)	1333 (0.8869)	1340 (0.8916)
	HP	72 (0.0479)	677 (0.4504)	1340 (0.8916)	1352 (0.8995)
	CL	289 (0.1923)	986 (0.6560)	1360 (0.9049)	**1387 (0.9228)**
	UBERON	488 (0.3247)	**1133 (0.7538)**	1358 (0.9035)	1371 (0.9122)
	GO+MP+HP+CL+UBERON	73 (0.0486)	728 (0.4844)	910 (0.6055)	1059 (0.7046)
SV pathogenicity prediction/ranking	StrVCTVRE	38 (0.0252)	620 (0.4125)	1022 (0.6799)	1343 (0.8935)
	AnnotSV	**930 (0.6188)**	930 (0.6188)	930 (0.6188)	**1493 (0.9933)**

Table 5: Summary of the evaluation for predicting causative variants in the synthetic whole genome benchmark dataset derived from dbVar, using 70% of the phenotypes.

		Synthetic dataset excluding 50 randomly of the phenotypes			
Aggregation method for genes within CNV	**PredCNV Models**	**Recall@1**	**Recall@10**	**Recall@30**	**Recall@50**
Maximum score	GO	27 (0.018)	538 (0.358)	827 (0.5502)	957 (0.6367)
	MP	70 (0.0466)	983 (0.654)	1137 (0.7565)	1207 (0.8031)
	HP	333 (0.2216)	1092 (0.7265)	**1370 (0.9115)**	1386 (0.9222)
	CL	27 (0.018)	538 (0.358)	828 (0.5509)	956 (0.6361)
	UBERON	74 (0.0492)	530 (0.3526)	623 (0.4145)	715 (0.4757)
	GO+MP+HP+CL+UBERON	171 (0.1138)	838 (0.5576)	936 (0.6228)	1049 (0.6979)
Average score	GO	235 (0.1564)	992 (0.6600)	1331 (0.8856)	1343 (0.8935)
	MP	27 (0.018)	915 (0.6088)	1331 (0.8856)	1339 (0.8909)
	HP	76 (0.0506)	675 (0.4491)	1333 (0.8869)	1352 (0.8995)
	CL	288 (0.1916)	987 (0.6567)	1360 (0.9049)	1390 (0.9248)
	UBERON	486 (0.3234)	**1133 (0.7538)**	1358 (0.9035)	1372 (0.9128)
	GO+MP+HP+CL+UBERON	70 (0.0466)	725 (0.4824)	910 (0.6055)	1074 (0.7146)
SV pathogenicity prediction/ranking	StrVCTVRE	38 (0.0252)	620 (0.4125)	1022 (0.6799)	1343 (0.8935)
	AnnotSV	**930 (0.6188)**	930 (0.6188)	930 (0.6188)	**1493 (0.9933)**

Table 6: Summary of the evaluation for predicting causative variants in the synthetic whole genome benchmark dataset derived from dbVar, using 50% of the phenotypes.

	Rank using maximum scores	**Rank using the average score**
GO	9	9
MP	7	10
HP	2	2
CL	9	12
UBERON	9	7
GO+MP+HP+CL+UBERON	**1**	5
StrVCTVRE	4	
AnnotSV	**1** (pr:0.090)	

Table 7: Ranking of disease-associated variant in a Saudi family using different DeepSVP models, and other methods. The ranks of the DeepSVP models are determined based on ranking a total of 47 variants. StrVCTVRE ranks only 6 out of 47 variants. AnnotSV predicted 11 variants as pathogenic out of 47 variants and we include the precision in parentheses.

Correlation between features

To test the redundancy of different scores of structural variants corresponding to the susceptibility of the phenotypes for the disease, we analyze the features by evaluating the pairwise correlation between them (Figure 5). According to their correlation coefficients, the features were broadly clustered into four major groups using the maximum and average score. As expected, measures of the structure of variants such as SV length and the number of genes or promoters were highly correlated. We further assess the importance of the features using two methods; the first using Extremely Randomized Trees ensemble learning Classifier (Extra Trees Classifier) (Figure 3), and the second using the Shapiro-Wilk algorithm [1] (Figure 4). Both methods explore the linear relationships across features; Extra Trees Classifier aggregates the results of multiple decision trees and outputs the ranked features based on the information gain, while Shapiro-Wilk assesses the normality of the distribution of examples with respect to the feature. We noticed that the features rank using both methods are similar; however, both do not capture potential nonlinear relationships among features, so we included all the ranked features in our model.

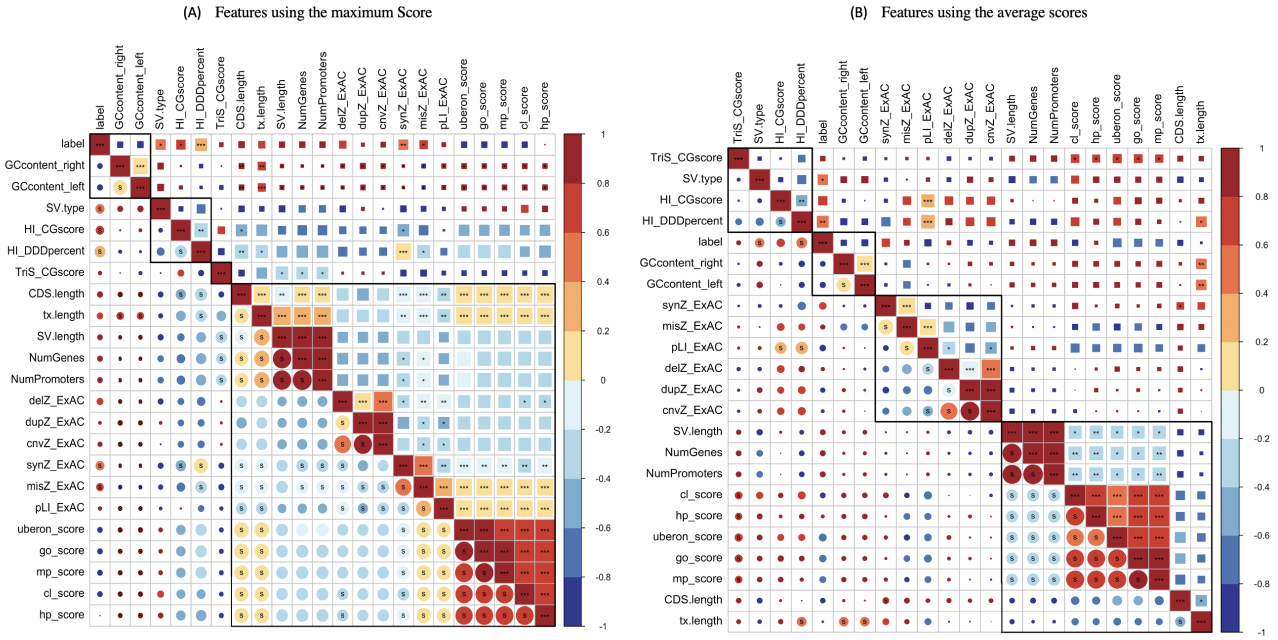

Figure 5: Correlation between the combined model features generated with *corrplot* package (version 0.84) in R, using *corrplot* function. 22 features of $42,202$ SVs, (8 features for the variant, and 14 derived from the genes overlapping the variant) were obtained from the disease phenotypes and AnnotSV using various databases. The pairwise correlation was computed on all the features. Figure **A** shows results using the maximum score, and Figure **B** the average score. The features are ordered, and different clusters are highlighted based on the hierarchical clustering. Significant correlations $(P < 0.05)$ are indicated by a letter 's' in the lower triangle. The color and size of circles represent the correlation strength (correlation coefficient). Statistical significance is indicated with asterisks (*P <0.05;** P <0.01;*** P <0.001).

References

[1] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.